Low-coverage whole-genome sequencing in livestock species for individual traceability and parentage testing

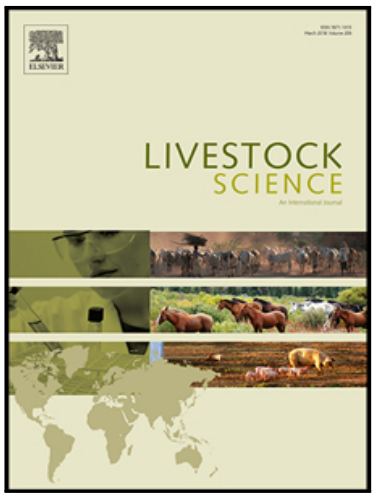
Casellas, Joaquim; de Hijas-Villalba, Melani Martín; Vázquez-Gómez, Marta; Id-Lahoucine, Samir

Published in:
Livestock Science

DOI:
10.1016/j.livsci.2021.104629

Print publication: 01/09/2021

Document Version
Version created as part of publication process; publisher's layout; not normally made publicly available



Low-coverage whole-genome sequencing in livestock species for individual traceability and parentage testing

Joaquim Casellas , Melani Martín de Hijas-Villalba , Marta Vázquez-Gómez , Samir Id-Lahoucine

PII: S1871-1413(21)00237-7
DOI: https://doi.org/10.1016/j.livsci.2021.104629
Reference: LIVSCI 104629

To appear in: *Livestock Science*

Received date: 24 March 2021
Revised date: 4 July 2021
Accepted date: 13 July 2021

Please cite this article as: Joaquim Casellas , Melani Martín de Hijas-Villalba , Marta Vázquez-Gómez , Samir Id-Lahoucine , Low-coverage whole-genome sequencing in livestock species for individual traceability and parentage testing, *Livestock Science* (2021), doi: https://doi.org/10.1016/j.livsci.2021.104629

Low-coverage whole-genome sequencing in livestock species for individual traceability and parentage testing

Joaquim Casellas[a,*], Melani Martín de Hijas-Villalba[a], Marta Vázquez-Gómez[a], Samir Id-Lahoucine[b]

[a]*Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain*

[b]*Animal and Veterinary Science Group, Scotland's Rural College, Edinburgh EH9 3JG, United Kingdom*

*Corresponding author.

E-mail address: joaquim.casellas@uab.cat (J. Casellas)

Highlights

- Traceability and paternity tests can adapt to low-coverage whole-genome sequencing data
- Testing performance depended on sequencing error rate and genotype frequencies
- Uncertainly had greater impact on false negatives than false positives
- 0.05× coverage sufficed to guarantee greater-than-99% success during testing

ABSTRACT

Procedures for genetic traceability of animal products and parentage testing mainly focus on microsatellites or SNPs panels. Nevertheless, current availability of high-throughput sequencing technologies must be considered as an appealing alternative. This research focused on the evaluation of low-coverage whole-genome sequencing for traceability and paternity testing purposes, within a context of evidential statistics. Analyses were performed on a simulation basis and assumed individuals with 30 100-Mb/100-cM chromosome pairs and ~1,000,000 polymorphic SNPs per chromosome. Ten independent populations were simulated under recombination and mutation with effective populations size 100 (generations 1 to 1,000), 10,000 (generation 1,001) and 25,000 (generation 1,002), and this last generation was retained for analytical purposes. Appropriate both traceability and paternity tests were developed and evaluated on different high-throughput sequencing scenarios accounting for genome coverage depth (0.01×, 0.05×, 0.1× and 0.5×), length of base-pair reads (100, 1,000 and 10,000 bp), and sequencing error rate (0%, 1% and 10%). Assuming true sequencing error rates and genotypic frequencies, 0.05× genome coverage depth guaranteed 100% sensitivity and specificity for traceability and paternity tests (n = 1,000). Same results were obtained when sequencing error rate was arbitrarily set to 0, or the the maximum value assumed during simulation (i.e., 1%). In a similar way, uncertainly about genotypic frecuencies did not impair sensitivity under 0.05× genome coverage, although it reduced specificity for paternity tests up to 85.2%. These results

highlighted low-coverage whole-genome sequencing as a promising tool for the livestock and food industry with both technological and (maybe) economic advantages.

Keywords: Evidential statistics, Paternity, Sequencing, Simulation, Traceability

1. Introduction

Neutral genetic markers have been widely used for both traceability (Arana et al., 2002; Vázquez et al., 2004) and parentage testing (Heaton et al., 2014) in livestock populations. Traceability aims to maintain credible custody of identification for animals or animal products through various steps within the production and food chain (McKean, 2001), and is becoming more demanding by consumers and producers (Qian et al., 2020). On the other hand, parentage testing enables to identify similar inheritance patterns between related individuals (Jamieson, 1965), and has a deep impact on breeding programs (Banos et al., 2001), where a moderate proportion of misidentified progeny can be anticipated (Geldermann et al., 1986; Visscher et al., 2002; Weller et al., 2004). Both approaches have relevant legal uses for animal forensic determinations (Kanthaswamy, 2015) or pedigree certification regarding livestock breed societies.

Genetic traceability and parentage testing rely on the fact that DNA is enormously variable among individuals despite the simple genetic mechanisms ruled by Mendel's laws of inheritance from parents to offspring. Moreover, DNA is present in every cell of the organism, does not change during animal life, and is stable to different treatments of processed food (Dalvit et al., 2007). Current procedures for genetic traceability and parentage testing mainly focus on microsatellites or SNPs (Heaton et al., 2002), where standardized panels have already been defined to harmonize procedures worldwide (https://www.isag.us/committees.asp, accessed March 18th, 2021).

Nevertheless, current advances in high-throughput technologies move towards partial or whole-genome sequencing procedures where closed SNP panels would be likely to have no future for further purposes. This requires additional endeavors to elucidate the usefulness of sequencing data, mainly when low-coverage approaches are considered due to economic limitations. Although Zan et al. (2019) suggested that very low-coverage (<0.5×) sequencing data could be informative enough for inferring outbred founder genotypes under an F_2 design, little is known about their applicability in commercial populations of livestock.

This research focused on three main objectives, (1) the development of both traceability and paternity tests for low-coverage sequencing data within the context of evidential inference (Bickel, 2012), (2) the validation of low-coverage sequencing for traceability and paternity testing in commercial livestock populations under full knowledge of population (*i.e.*, allele or genotype frequencies) and sequencing parameters (*i.e.*, error rates), and (3) the evaluation of the impact of uncertainly about population and sequencing parameters on traceability and paternity tests for low-coverage sequencing data.

2. Materials and methods

Animal Care and Use Committee approval was not obtained for this study because analyses were performed on simulated data sets. Neither real

animals nor biological tissues from alive animals were involved in this research.

2.1. Genome and population simulation process

This research simulated an unspecific mammalian livestock population. We took as a starting point a 100-Mb/100-cM chromosome with 5,000,000 biallelic SNPs (one SNP each 20 base pairs and 2×10^{-5} cM), and the whole genome consisted of 30 chromosome pairs. This generated a standard 3 Gb genome (Pérez-Enciso et al., 2015) with the same number of chromosomes as cattle and goat, and within the range of other livestock species such as pig (19 pairs), sheep (27 pairs) and horse (32 pairs). The starting number of SNPs was assumed to guarantee more than 30,000,000 polymorphic SNPs at the end of the simulation process (see below), as reported by Daetwyler et al. (2014) in cattle.

Populations started from a founder generation with 100 individuals that were heterozygous throughout the whole genome. They evolved during 1,000 non-overlapping generations under random mating and effective population size 100. Linkage disequilibrium between adjacent loci was generated based on Kosambi's mapping function (Kosambi, 1944), and a mutation rate of 2.5×10^{-3} per SNP was applied until generation 980 (Meuwissen et al., 2001), switching the allele state from A to B, or vice versa. From generation 981 on, the mutation rate switched to 2.5×10^{-8} (Hickey and Gorjanc, 2012). Only those populations with 1,000,000 ± 10% (*i.e.*, 900,000 to 1,100,000) polymorphic (MAF > 0) SNPs per chromosome in generation 1,000 were retained for further analyses.

Populations expanded to 10,000 individuals in generation 1,001 (1,000

sires and 9,000 dams), and 25,000 in generation 1,002. A total of 10 independent populations were simulated.

2.2. Sequencing and alignment simulation process

The number of reads per chromosome was defined as

$$C \times (100 \times 10^6) / L,$$

where 100×10^6 was the assumed chromosome length in bp, C was the expected genome coverage, and L was the average read length in base-pairs. The length of each read was sampled from a normal distribution with mean L and standard deviation L/10 to account for variability on DNA sequencing products. Moreover, each read was placed at random, both in the genome and chromosome phase. Following Fox et al. (2014) and Pfeiffer et al. (2018), an error rate between 10^{-5} and 10^{-2} was randomly assigned to each polymorphic SNP. The same error rate applied to both alleles. Only the number of reads for each allele was stored for further analyses.

2.3. Evidential testing for single-individual traceability

This research relied on evidential inference (Edwards, 1972) as a way to compare two competing hypotheses (*i.e.*, models). This approach relies on the likelihood function as the structure that contains all evidence from the data relevant to the statistical model (Birnbaum, 1962), and compares hypotheses by calculating the ratio of their likelihood functions (Hacking, 1965). Within this context, an upper-than-1 likelihood ratio favors the numerator model whereas a

lower-than-1 likelihood ratio suggests the superiority of the denominator model, although a minimum likelihood ratio of 32 (or 1/32) is typically used in the evidential literature (Blume, 2002; Royall, 1997), or even as high as 1,000, often used in genome-wide linkage studies (Morton, 1998).

Traceability in the livestock industry can be defined as the ability to identify animals or animals products through various steps within the food chain from the farm to the retailer (McKean, 2001). Within this context, the analysis of genetic polymorphisms must be viewed as a key tool to verify the match between two independent samples.

Take as a starting point a $n \times 2$ matrix ($\mathbf{S}$) to summarize sequence data from n polymorphic and biallelic sites of the genome. Once sorted by chromosome and nucleotide within the chromosome, each column stores the number of reads for alleles A and B, respectively. The analysis of genetic traceability relied on two different samples ($\mathbf{S}_p$ and $\mathbf{S}_q$), and two competing hypotheses, *i.e.*, H_0: samples belong to the same individual ($p = q$), and H_1: both samples belong to different individuals ($p \neq q$). They can be tested through their likelihood ratio (Edwards, 1972) as follows,

$$LR(H_0, H_1 \mid \mathbf{S}_p, \mathbf{S}_q) = p(\mathbf{S}_p, \mathbf{S}_q \mid H_0) \,/\, p(\mathbf{S}_p, \mathbf{S}_q \mid H_1),$$

where $p(\mathbf{S}_p, \mathbf{S}_q \mid H_k)$ was the joint probability of obtaining data $\mathbf{S}_p$ and $\mathbf{S}_q$ under hypothesis H_k. Under the H_0 hypothesis, the likelihood must expand to

$$p(\mathbf{S}_p, \mathbf{S}_q \mid H_0) = \Pi_{i=1,n}\, p(\mathbf{s}_{p,i} \mid g_{p,i}, \varepsilon_i)\, p(\mathbf{s}_{q,i} \mid g_{p,i}, \varepsilon_i)\, p(g_{p,i})$$

where $\mathbf{s}_{p,i}$ was the ith row of $\mathbf{S}_p$, $g_{p,i}$ was the genotype (*i.e.*, AA, AB or BB) of the pth individual for the ith polymorphic site, and ε_i was the sequencing error rate

for the *i*th polymorphic site (we assume homogeneous error rates among alleles). Given that *p* and *q* were assumed to be the same individual and the parametric space accounted for three genotypes (p(AA) + p(AB) + p(BB) = 1), the likelihood became

$$p(\mathbf{S}_p, \mathbf{S}_q \mid H_0) = \Pi_{i=1,n}[\Sigma_{\alpha=AA,AB,BB}\, p(\mathbf{s}_{p,i} \mid g_{p,i} = \alpha\,,\, \varepsilon_i)\, p(\mathbf{s}_{q,i} \mid g_{q,i} = \alpha\,,\, \varepsilon_i)\, p(g_{p,i} = g_{q,i} = \alpha)]$$

Now, assume *a* reads for allele A and *b* reads for allele B in $\mathbf{s}_{p,i}$. The following conditional probabilities can be straightforwardly calculated as binomial processes with trials, successes and success probability sequentially noted between parentheses,

$$p(\mathbf{s}_{p,i} \mid g_{p,i} = AA\,,\, \varepsilon_i) = Binomial(a+b,\, a,\, 1 - \varepsilon_i)$$

$$p(\mathbf{s}_{p,i} \mid g_{p,i} = AB\,,\, \varepsilon_i) = Binomial(a+b,\, a,\, 0.5)$$

$$p(\mathbf{s}_{p,i} \mid g_{p,i} = BB\,,\, \varepsilon_i) = Binomial(a+b,\, a,\, \varepsilon_i),$$

Finally, the probability of each genotype depends on its frequency in the source population.

The same development can be applied to the alternative hypothesis where

$$p(\mathbf{S}_p, \mathbf{S}_q \mid H_1) = \Pi_{i=1,n}\, p(\mathbf{s}_{p,i} \mid g_{p,i}, \varepsilon_i)\, p(g_{p,i})\, p(\mathbf{s}_{q,i} \mid g_{q,i}, \varepsilon_i)\, p(g_{q,i}),$$

and *p* and *q* were assumed different and unrelated individuals from the same population. Once accounted for all three possible genotypes, the previous expression expanded to

$$p(\mathbf{S}_p, \mathbf{S}_q \mid H_1) = \Pi_{i=1,n}\, \{[\Sigma_{\alpha=AA,AB,BB}\, p(\mathbf{s}_{p,i} \mid g_{p,i} = \alpha\,,\, \varepsilon_i)\, p(g_{p,i} = \alpha)]$$

$$\times\, [\Sigma_{\beta=AA,AB,BB}\, p(\mathbf{s}_{q,i} \mid g_{q,i} = \beta\,,\, \varepsilon_i)\, p(g_{q,i} = \beta)]\}.$$

2.4. Testing for parentage

Parentage testing relies on the use of biological markers to identify similar inheritance patterns between related individuals and traces back to the 1960s where blood typing was used as a regular part of some cattle breeding programs (Stormont, 1967). As seen with most domestic species, the typical animal parentage case includes a dam, offspring, and one or more alleged sires. The identity of the dam uses to be fairly certain, whereas the true sire must be identified from a set of *m* males. Our analytical approach will rely on this scenario, although it can be straightforwardly generalized to test the other sex (*i.e.*, dam).

Paternity testing relied on data samples from the offspring ($\mathbf{S}_o$), its dam ($\mathbf{S}_d$), and an alleged sire ($\mathbf{S}_s$). The testing process started with the definition of the null hypothesis such as $H_{0,i}$: both *s* and *d* were parents of *o*. Within this context, the joint likelihood of $\mathbf{S}_o$, $\mathbf{S}_d$ and $\mathbf{S}_s$ was written as

$$p(\mathbf{S}_o, \mathbf{S}_d, \mathbf{S}_s \mid H_0) = \Pi_{i=1,n}\{p(\mathbf{s}_{o,i} \mid g_{o,i}, \varepsilon_i)\, p(g_{o,i} \mid g_{d,i}, g_{s,i})\, p(\mathbf{s}_{d,i} \mid g_{d,i}, \varepsilon_i)\, p(g_{d,i}) \times p(\mathbf{s}_{s,i} \mid g_{s,i}, \varepsilon_i)\, p(g_{s,i})\},$$

where $\mathbf{s}_{o,i}$ was the *i*th row of $\mathbf{S}_o$, $g_{o,i}$ was the genotype of the *o*th individual in the *i*th polymorphic site, and ε_i was the sequencing error rate for the *i*th polymorphic site (we assume homogeneous error rates among alleles). As for traceability tests, previous likelihood expanded to account for biallelic genetic markers,

$$p(\mathbf{S}_o, \mathbf{S}_d, \mathbf{S}_s \mid H_0) = \Pi_{i=1,n}\{\Sigma_{\alpha=AA,AB,BB}\, p(\mathbf{s}_{o,i} \mid g_{o,i} = \alpha, \varepsilon_i)$$

$$\times [\Sigma_{\beta=AA,AB,BB} \Sigma_{\gamma=AA,AB,BB} p(g_{o,i} = \alpha \mid g_{d,i} = \beta , g_{s,i} = \gamma) p(\mathbf{s}_{d,i} \mid g_{d,i} = \beta , \varepsilon_i) p(g_{d,i} = \beta)$$

$$\times p(\mathbf{s}_{s,i} \mid g_{s,i} = \gamma , \varepsilon_i) p(g_{s,1} = \gamma)]\},$$

where $p(\mathbf{s}_{o,i} \mid g_{o,i} = \alpha , \varepsilon_i)$, $p(\mathbf{s}_{d,i} \mid g_{d,i} = \beta , \varepsilon_i)$ and $p(\mathbf{s}_{s,i} \mid g_{s,i} = \gamma, \varepsilon_i)$ were binomial probabilities, $p(g_{d,i} = \beta)$ and $p(g_{s,i} = \gamma)$ were genotypic frequencies in the parental population, and $p(g_{o,i} = \alpha \mid g_{d,i} = \beta , g_{s,i} = \gamma)$ was the conditional probability of the offspring's genotype depending on parents' genotype (Table 1). It is important to note that previous expression can also be applied when lacking of sequencing data from the dam as follows,

$$p(\mathbf{S}_o, \mathbf{S}_s \mid H_0) = \Pi_{i=1,n}\{p(\mathbf{s}_{o,i} \mid g_{o,i}, \varepsilon_i) p(g_{o,i} \mid g_{s,i}) p(\mathbf{s}_{s,i} \mid g_{s,i} , \varepsilon_i) p(g_{s,i})\},$$

$$p(\mathbf{S}_o, \mathbf{S}_s \mid H_0) = \Pi_{i=1,n} \{\Sigma_{\alpha=AA,AB,BB} p(\mathbf{s}_{o,i} \mid g_{o,i} = \alpha , \varepsilon_i)$$

$$\times [\Sigma_{\gamma=AA,AB,BB} p(g_{o,i} = \alpha \mid g_{s,i} = \gamma) p(\mathbf{s}_{s,i} \mid g_{s,i} = \gamma , \varepsilon_i) p(g_{s,1,i} = \gamma)]\},$$

where $p(g_{o,i} = \alpha \mid g_{s,i} = \gamma)$ can be obtained from Table 2.

On the other hand, the alternative hypothesis could be defined on the following rationale, H_1: only *d* was parent of *o*, whereas *s* was unrelated to *o* and sampled from the same population. The likelihood expands to

$$p(\mathbf{S}_o, \mathbf{S}_d, \mathbf{S}_s \mid H_1) = \Pi_{i=1,n} \{p(\mathbf{s}_{o,i} \mid g_{o,i}, \varepsilon_i) p(g_{o,i} \mid g_{d,i})$$

$$\times p(\mathbf{s}_{d,i} \mid g_{d,i}, \varepsilon_i) p(g_{d,i})\} p(\mathbf{s}_{s,i} \mid g_{s,i} , \varepsilon_i) p(g_{s,i}),$$

and

$$p(\mathbf{S}_o, \mathbf{S}_d, \mathbf{S}_s \mid H_1) = \Pi_{i=1,n} \{\Sigma_{\alpha=AA,AB,BB} p(\mathbf{s}_{o,i} \mid g_{o,i} = \alpha , \varepsilon_i)$$

$$\times [\Sigma_{\beta=AA,AB,BB} p(g_{o,i} = \alpha \mid g_{d,i} = \beta) p(\mathbf{s}_{d,i} \mid g_{d,i} = \beta , \varepsilon_i) p(g_{d,i} = \beta)]\}$$

$$\times \Pi_{i=1,n} [\Sigma_{\gamma=AA,AB,BB} p(\mathbf{s}_{s,i} \mid g_{s,i} = \gamma , \varepsilon_i) p(g_{s,i} = \gamma)].$$

where $p(g_{o,i} = \alpha \mid g_{d,i} = \beta)$ can be found in Table 2. As for previous hypothesis, it was not mandatory to account for dam sequencing data if missing,

$$p(\mathbf{S}_o, \mathbf{S}_s \mid H_1) = \Pi_{i=1,n}\, p(\mathbf{s}_{o,i} \mid g_{o,i}, \varepsilon_i)\, p(g_{o,i})\, p(\mathbf{s}_{s,i} \mid g_{s,i}, \varepsilon_i)\, p(g_{s,i}),$$

$$p(\mathbf{S}_o, \mathbf{S}_s \mid H_1) = \Pi_{i=1,n} \Sigma_{\alpha=AA,AB,BB}\, p(\mathbf{s}_{o,i} \mid g_{o,i} = \alpha, \varepsilon_i)\, p(g_{o,i} = \alpha)$$

$$\times \Sigma_{\gamma=AA,AB,BB}\, p(\mathbf{s}_{s,i} \mid g_{s,i} = \gamma, \varepsilon_i)\, p(g_{s,i} = \gamma).$$

2.5. Uncertainly about population and sequencing parameters

Single individual traceability and parentage testing were evaluated under different scenarios accounting for 0.01×, 0.05×, 0.1× and 0.5× depth of genome coverage, with 100, 1,000 and 10,000 base-pair reads. Those read lengths were chosen to illustrate test performance under currently available sequencing platforms (Besser et al., 2018).

As noted above, both traceability and parentage tests relied on two structural parameters, within-SNP sequencing error rate (ε_i) and genotyping frequencies. The first mainly depends on the sequencing method and platform used (Fox et al., 2014) and uses to be estimated on an across-genome basis. Within this context, we compared test performances under three across-SNP homogeneous sequencing error rates: 0%, 1% (the maximum sequencing error rate used for simulation of the sequencing process), and 10% (*i.e.*, ten times higher than the maximum sequencing error rate used for simulation of the sequencing process).

On the other hand, genotypic frequencies could be approximated by using sequence data generated for traceability and paternity tests. Nevertheless, the number of sequenced animals could be small and contribute high uncertainly to estimated genotypic frequencies. To account for this

uncertaintly, the variance of the estimated A allele frequency (π) can be calculated as (Cockerham, 1969)

$$V(\pi) = [\pi (1 - \pi)] / 2\lambda$$

where λ was the number of sampled individuals. We compared λ = 5, 10 and 100, and sampled the A allele frequency (π*) for each SNP from a truncated (0 to 1) normal distribution with mean π and variance V(π). Genotypic frequencies were obtained assuming Hardy-Weinberg equilibrium (Hardy, 1908).

3. Results

3.1. Simulated genomic data

After 1,000 non-overlapping generations, random mating and effective population size 100, we retained ten populations with 29,195,811 to 30,660,474 polymorphic SNPs. Allele frequencies widely distributed along with the parametric space, as shown in Fig. 1, and a remarkable percentage of SNPs had minimum allele frequency (MAF) below 0.05. Although this varied among chromosomes, between 36.7% and 51.9% of SNPs had MAF < 0.05. All these 10 simulated populations contributed equally to the subsequent analyses.

After sequencing 10,000 individuals, the maximum number of reads per polymorphic SNP was 3 (0.01× genome coverage), 4 (0.05×), 5 (0.1×) and 7 (0.5×). Nevertheless, between 76.2% (0.5× genome coverage) and 99.5% (0.01× genome coverage) of them had a single read, as shown in Fig. 2. The percentage of polymorphic SNPs with two reads increased with genome

coverage, from 0.5% (0.01×) to 19.1% (0.5×), and a similar trend with smaller percentages was revealed for larger numbers of reads. Moreover, those percentages showed small variability across individuals, this uncertainly even reducing for smaller read length (Fig. 2). The same pattern was revealed when checking for shared SNPs among pairs of sequenced individuals. The longer the read length was, the wider the dispersion of the number of shared SNPs (Fig. 3). From the total of ~30,000,000 polymorphic SNPs, the average number of shared polymorphic SNPs decreased from 3,355.6 ± 4.7 (100 base-pair read length) to 3,093.2 ± 28.5 (10,000 base-pair read length). For SNPs with MAF ≥ 0.05, similar trends were observed, from 1,748.4 ± 3.3 (100 base-pair read length) to 1,586.9 ± 17.2 (10,000 base-pair read length). Within this context, subsequent results were reported based on the most uncertain (*i.e.*, increased variability for the number of reads and shared SNPs) and less informative (*i.e.*, reduced number of shared SNPs) scenario, this accounting for sequencing by 10,000 base-pair reads.

3.2. Traceability and parentage testing

As anticipated, the number of shared polymorphic SNPs among two unrelated individuals quickly increased with genome coverage (Table 3). This generated a fast growth in terms of available information for traceability and paternity tests, as evidenced by the likelihood ratios provided in Fig. 4. Assuming true sequencing error rates and genotype frequencies, 100% of traceability tests favored the true hypothesis when genome coverage was 0.05×

or deeper. The only exceptions were detected for 0.01X genome coverage, where 0.7% of false positives and 0.04% of false negatives were reported (Fig. 4). The same pattern was revealed for paternity tests, they showing a 100% of true positive and true negatives under genome coverage 0.05X or deeper, and 1.1% (1.0%) of false positives and 0.5% (0.6%) of false negatives under 0.01X of genome coverage when the dam was known (unknown).

In order to test for a more realistic scenario, different homogeneous sequencing error rates were evaluated. As shown in Fig. 5, 0.05X coverage sufficed to avoid false positives and negatives under both traceability and paternity tests when sequencing error rate was arbitrarily set to 0% or the maximum rate used during sequencing simulation (*i.e.*, 1%). The only assumption that generated wrong results under 0.05X coverage was when the sequencing error rate was unrealistically assumed 10 times higher than the upper bound during sequencing (*i.e.*, 10%). In this case, 1.3% (traceability test), 24.5% (paternity test with known dam) and 22.8% (paternity test with an unknown dam) of false negatives were reported, whereas any test generated false positives. Higher genome coverage tested provided 100% of true positives and true negatives (results not shown), even under the assumption of 10% of the sequencing error rate.

The other parameter accounting for uncertainly during traceability and paternity testing was genotype frequencies. In this case, genotype frequencies were assumed under Hardy-Weinberg equilibrium and calculated from allele frequency with uncertainly as sampled from 5, 10, and 100 individuals. As

shown in Fig. 6, the smaller the uncertaintly for allele frequency was, the larger the match with results was obtained under true genotype frequencies. Nevertheless, 0.05× coverage sufficed to avoid both false positives and negatives in traceability tests, whatever the accuracy of allele frequencies. Paternity tests with known dam (unknown dam) revealed a similar pattern without false positives since 0.05× coverage, and 0.1% (0.1%), 0.6% (0.3%) and 10.2% (14.8%) of false negatives when the allele frequency was sampled with uncertainly as calculated from 100, 10 and 5 individuals. Deeper genome coverage provided 100% of true positives and true negatives at any uncertainly for allele frequency.

4. Discussion

Current procedures for traceability and paternity testing rely on SNPs where standardized panels have already been defined to harmonize procedures worldwide (Heaton et al., 2002). Although their reliability and statistical power fulfill the purpose for which they were created (Marshall et al., 1998), they depend on some dozens of a few hundreds of SNP genotypes, too few to be reused for other purposes like genome-wide association analyses (Klein et al., 2005; Gilly et al., 2019) or genomic evaluation (Meuwissen et al., 2001; Gorjanc et al., 2015, 2017). This is an important limitation because it drains the economic capacity of food chain industries and breed societies and precludes additional investments in genomic techniques. The current explosion in high-

throughput sequencing technologies (Bansal and Boucher, 2019) opens the door to more sustainable science where specificity and multiple-purpose data are not conflicting terms. Nevertheless, a first step is required to verify that low-coverage whole-genome sequencing data can efficiently address both traceability and paternity tests in order to fulfill current standards at a similar economic cost.

Theoretical approaches to test both traceability and paternity have been widely developed in scientific literature on the basis of complete genotypes (Goffaux et al., 2005; Martin et al., 2010; Marshall et al., 1998), whereas high-throughput sequencing technologies provide a variable number of random samples from each polymorphic site and require genotype-calling procedures to reach closed genotypes (Nielsen et al., 2011). Nevertheless, genotype-calling approaches show little agreement when compared under low-coverage sequencing data (Liu et al., 2013; Vens et al., 2009; Yu and Sun, 2013), where heterozygous genotypes cannot be adequately called with a single read (Brouard et al., 2017). Within this context, we omitted genotype-calling approaches in our traceability and paternity tests and focused on genotype probabilities within the context of appropriate likelihood functions. Although these procedures were partially implemented in some genotype-calling approaches (Li et al., 2008, 2009; Martin et al., 2010), they summarized to the most probable genotype instead of keeping uncertainly for further analyses. We kept uncertainly about genotypes along the whole calculation of the likelihood ratio in order to avoid arbitrary decisions when available information for each

polymorphic site was very small in tested individuals (Fig. 2).

Our tests relied on the likelihood principle, a statistical proposition that states that all the evidence in the data relevant to the statistical model is contained in the likelihood function (Birnbaum, 1962). Within this context, a likelihood ratio must be viewed as an objective measurement of the statistical evidence of one model against the other (Hacking, 1965), and establishes the foundations for the evidential statistics (Edwards, 1972) in contrast with frequentist and Bayesian statistics. This inferential approach relies on two basic conditions that are not completely fulfilled by frequentist and Bayesian inferences, objectivity (*i.e.*, the strength of evidence does not vary from one researcher to another) and interpretability (*i.e.*, the strength of evidence has the same practical interpretation for any sample size). The first condition rules out Bayes factors that depend on subjective or default priors (Bickel, 2012), and the second rules out the frequentist *p*-value that forces the same type-I error percentage at any sample size (Bickel, 2011). By contrast, the likelihood ratio satisfies both of the necessary conditions for a measure of the strength of statistical evidence. Within this context, the likelihoods used in our testing approaches had the same mathematical structure than the likelihoods we could construct within a frequentist scenario, as well as they are proportional to the joint posterior distributions with flat priors we could call in the Bayesian framework. The essential difference relies on the test itself and the assumptions carried out by the researcher. Within the context of evidential statistics, there are not additional assumptions apart from the statistical model itself and all the

hypotheses have the same consideration during the analytical process. Indeed, paternity tests with panels of genetic markers were previously proposed by Marshal et al., (1998), and evidential statistics have been growing attention in genetics and genomics research (Strug et al., 2010; Strug, 2018).

The performance of both traceability and paternity tests was outstanding as evidenced in Figs. 4 to 6. Under the unrealistic assumption of known sequencing error rates and genotype frequencies, 0.01× genome coverage sufficed to guarantee ≥99% of true positives and true negatives under traceability tests. In contrast, the minimum genome coverage for paternity tests must increase up to 0.05× genome coverage to reach the same rate of true positives and negatives. Nevertheless, our current method works with low-coverage sequencing data and less false paternity assignments than previous methods found in the scientific literature (Snyder-Mackler et al., 2016; Whalen et al., 2019). The method design for very low sequencing coverage data from fecal-derived DNA by Snyder-Mackler et al. (2016), which also performed peternity tests wit known or unknown dam, was not available to assign paternity below 0.17×. On the other hand, results for paternity analyses by Whalen et al. (2019) required greater coverage (0.4×) and larger amount of genetic markers (50,000) to reach 100% sensibility.

In order to evaluate those procedures under more realistic scenarios, different homogeneous error rates and accuracies for genotype frequencies were evaluated. In this case, the sequencing error rate had a mild impact on the performance of both traceability and paternity tests, and it only impaired their

results when an abnormally high sequencing error rate was assumed (*i.e.*, 10%). Indeed, results shown in Fig. 5 suggested that the assumption of a null sequencing error rate provided the most similar results to the ones obtained under true sequencing error rates, simplifying both analytical models and subsequent calculations. On the other hand, the impact of genotype frequencies was suggested as larger, where more accurate estimates were required to avoid false positives and negatives.

Statistical methodologies developed in this manuscript are ready to use for both the food chain industry and breed societies. In fact, they could also be useful for human studies. They do not need additional generalizations, as all required algorithms are detailed in the current manuscript. It is important to highlight that 0.05X genome coverage sufficed for traceability and paternity tests assuming null (or 1%) sequencing error rate and an accuracy for allele frequencies equal or higher to the ones obtained when sampling 10 individuals. This must be viewed as an outstanding result from technological, economic and scientific points of view. Moreover, the sequenceing data generated could have firther uses contributing more to sustainable science. The huge amount of information available (even under very-low coverage) can be exploited more in depth. Especially, with the structure of livestock species with dense family structures, large amounts of genomic data can accumulate across generations and years. This latter will open a new window of animal breeding purposes, as the availability of whole sequence for animal population may change the current animal breeding paradigm or even make a new revolution. Indeed, the

exploration of sequence data at massive volume may allow to make animal breeding selection decisions more accurate by taking benefit of massive genomic data (Knap et al., 2020). Thus, additional efforts to handle this new source of partial genomic data may be of special relevance for the livestock industry (Knap et al., 2020). Evenmore, an additional investment to increase the sequencing coverage until 2×, which is still considered low-coverage, could allow to enhance animal breeding. Between the possible options are the estimation of biological relatedness (Lipatov et al., 2015) and the imputation of the whole genome with high accuracy depending on the population size (Ros-Freixedes et al., 2020a, 2020b). This last step would be essential to implement whole-genome sequence data for genomic prediction and fine-mapping of causal variants.

5. Conclusions

Very low genome coverages in livestock species were enough to guarantee ≥99% of true positives and true negatives for traceability testing (from 0.01× coverage) and parentage testing (from 0.05× coverage). Even when 0.05× coverage sufficed for both tests, as genome coverage increased, the percentage of reads per polymorphic SNPs and the certainty of the estimate of its allele frequency increased, thus, reducing the errors in the tests. Moreover, the length of the reads affected the dispersion and number of shared SNPs among pairs of sequenced individuals.

Author statement
Joaquim Casellas: Conceptualization, Data Simulation, Formal analysis, Software, Writing – original draft; **Melani Martín de Hijas-Villalba:** Methodology, Writing – review & editing; **Marta Vázquez-Gómez:** Methodology, Writing – review & editing; **Samir Id-Lahoucine:** Methodology, Writing – review & editing.
Conflict of interest statement

There are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Acknowledgments

This research was partially supported by the grant CGL2016-80155-R (Spain's *Ministerio de Economía y Competitividad*, Madrid, Spain) and a fellowship granted to M. Martín de Hijas-Villalba (grant BES-2017-080596; Spain's *Ministerio de Economía y Competitividad*, Madrid, Spain).

Conflict of interest statement

Authors declare no conflict of interest.

References

Arana, A., Soret, B., Lasa, I., Alfonso, L., 2002. Meat traceability using DNA

markers: application to the beef industry. Meat Sci. 61, 367-373.

Banos, G., Wiggans, G.R., Powell, R.L., 2001. Impact of paternity errors in cow identification on genetic evaluations and international comparisons. J. Dairy Sci. 84, 2523-2529.

Bansal, V., Boucher, C., 2019. Sequencing technologies and analyses: where have we been and where are we going? Science 18, 37-41.

Besser, J., Carleton, H.A., Gerner-Smidt, P., Lindsey, R.L., Trees, E., 2018. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clin. Microbiol. Infect. 24, 335-341.

Bickel, D.R., 2011. A predictive approach to measuring the strength of statistical evidence for single and multiple comparisons. Can. J. Stat. 39, 610-631.

Bickel, D.R., 2012. The strength of statistical evidence for composite hypotheses: inference to the best explanation. Stat. Sin. 22, 1147-1198.

Birnbaum, A., 1962. On the foundations of statistical inference. J. Am. Stat. Assoc. 57, 269-306.

Blume, J.D., 2002. Tutorial in biostatistics: likelihood methods for measuring statistical evidence. Stat. Med. 21, 2563-2599.

Brouard, J.-S., Boyle, B., Ibeagha-Awemu, E.M., Bissonnette, N., 2017. Low-depth genotyping-by-sequencing (GBS) in a bovine population: strategies to maximize the selection of high quality gentypes and the accuracy of imputation. BMC Genet. 18, 32.

Cockerham, C.C., 1969. Variance of gene frequencies. Evol. 23, 72-84.

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C., Esquerré, D., Bouchez, O., Rossignol, M.-N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P.J., Coote, D., Chamberlain, A.J., Anderson, C., VanTassell, C.P., Hulsegge, I., Goddard, M.E., Guldbrandtsen, B., Lund, M.S., Veerkamp, R.F., Boichard, D., Fries, R., Hayes, B.J., 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46, 858-867.

Dalvit, C., De Marchi, M., Cassandro, M., 2007. Genetic traceability of livestock products: A review. Meat Sci. 77, 437-449.

Edwards, A.W.F., 1972. Likelihood. Cambridge University Press, Cambridge, United Kingdom.

Fox, E.J., Reid-Bayliss, K.S., Emond, M.J., Loeb, L.A., 2014. Accuracy of next generation sequencing platforms. Next Gen. Seq. Appl. 1, 1000106.

Geldermann, H., Piper, U., Weber, W.E., 1986. Effect of misidentification on the estimation of breeding value and heritability in cattle. J. Anim. Sci. 63, 1759-1768.

Gilly, A., Southam, L., Suveges, D., Kuchenbaecker, K., Moore, R., Melloni, G.E.M., Hatzikotoulas, K., Farmaki, A.-E., Ritchie, G., Schwartzentruber, J., Danecek, P., Kilian, B., Pollard, M.O., Ge, X., Tsafantakis, E., Dedoussis, G., Zeggini, E., 2019. Very low-depth whole-genome sequencing in complex trait association studies. Bioinformatics 35, 2555-2561.

Goffaux, F., China, B., Dams, L., Clinquart, A., Daube, G., 2005. Development of a genetic traceability test in pig based on single nucleotide polymorphism detection. Forensic Sci. Int. 1651, 239-247.

Gorjanc, G., Cleveland, M.A., Houston, R.D., Hickey, J.M., 2015. Potential of genotyping-by-sequencing for genomic selection in livestock populations. Genet. Sel. Evol. 47, 12.

Gorjanc, G., Dumasy, J.-F., Gonen, S., Gaynor, R.C., Antolin, R., Hickey, J., 2017. Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregation populations. Crop Sci. 57, 1404-1420.

Hacking, I., 1965. Logic of statistical inference. Cambridge University Press, Cambridge, United Kingdom.

Hardy, G.H., 1908. Mendelian proportions in a mixed population. Science 28, 49-50.

Heaton, M.P., Harhay, G.P., Bennett, G.L., Stone, R.T., Grosse, W.M., Casas, E., Keele, J.W., Smith, T.P., Chitko-McKown, C.G., Laegreid, W.W., 2002. Selection and use of SNP markers for animal identification and paternity analysis in US beef cattle. Mamm. Genome 13, 272-281.

Heaton, M.P., Leymaster, K.A., Kalbfleisch, T.S., Kijas, J.W., Clarke, S.M., McEwan, J., Maddox, J.F., Basnayake, V., Petrik, D.T., Simpson, B., Smith, T.P.L., Chitko-McKown, C.G., Int. Sheep. Genom. Cons., 2014. SNPs for parentage testing and traceability in globally diverse breeds of sheep. PLoS One 9, e94851.

Hickey, J.M., Gorjanc, G., 2012. Simulated data for genomic selection and genome-wide association studies using a combination of coalescent gene drop methods. G3 (Bethesda) 2, 425-427.

Jamieson, A., 1965. The genetics of transferrins in cattle. Heredity 20, 419-441.

Kanthaswamy, S., 2015. Domestic animal forensic genetics – Biological evidence, genetic markers, analytical approaches and challenges. Anim. Genet. 46, 473-484.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C., Hoh, J., 2005. Complement factor H polymorphism in age-related macular degeneration. Science 308, 385-389.

Knap, P.W., 2020. The scientific development that we need in the animal breeding industry. J. Anim. Breed. Genet. 137, 343-344.

Kosambi, D., 1944. The estimation of map distances from recombination values. Ann. Eugen. 12, 172-175.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., Wang, J., 2009. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25, 1966-1967.

Li, H., Ruan, J., Durbin, R.M., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18, 1851-1858.

Lipatov, M., Sanjeev, K., Patro, R., Veeramah, K.R., 2015. Maximum Likelihood

Estimation of Biological Relatedness from Low Coverage Sequencing Data. bioRxiv 023374.

Liu, X., Han, S., Wang, Z., Gelernter, J., Yang, B.-Z., 2013. Variant callers for next-generation sequencing data: a comparison study. PLoS One 8, e75619.

Marshall, T.C., Slate, J., Kruuk, L.E.B., Pemberton, J.M., 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol. Ecol. 7, 639-655.

Martin, E.R., Kinnamon, D.D., Schmidt, M.A., Powell, E.H., Zuchner, S., Morris, R.W., 2010. SeqEM: an adaptive genotyping-calling approach for next-generation sequencing studies. Bioinformatics 26, 2803-2810.

McKean, J.D., 2001. The importance of traceability for public health and consumer protection. Rev. Sci. Tech. 20, 363-371.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819-1829.

Morton, N., 1998. Significance levels in complex inheritance. Am. J. Hum. Genet. 62, 690–697.

Nielsen, R., Paul, J.S., Albrechtsen, A., Song, Y.S., 2011. Genotype and SNP from next-generation sequencing data. Nat. Rev. Genet. 12, 443-451.

Pérez-Enciso, M., Rincón, J.C., Legarra, A., 2015. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet. Sel. Evol. 47, 43.

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J.L., Mayer, G., 2018. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Sci. Rep. 8, 10950.

Qian, J., Ruíz-García, L., Fan, B., Robla-Villalba, J.I., McCarthy, U., Zhang, B, Yu, Q., Wu, W., 2020. Food traceability system for governmental, corporate, and consumer perspectives in the European Union and China: A comparative review. Trends Food Sci. Technol. 99, 402-412.

Ros-Freixedes, R., Whalen, A., Chen, C.-Y., Gorjanc, G., Herring, W.O., Mileham, A.J., Hickey, J.M., 2020a. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. Genet. Sel. Evol. 52, 17.

Ros-Freixedes, R., Whalen, A., Gorjanc, G., Mileham, A.J., Hickey, J.M., 2020b. Evaluation of sequencing strategies for whole-genome imputation with hybrid peeling. Genet. Sel. Evol. 52, 18.

Royall, R.M., 1997. Statistical Evidence: A Likelihood Paradigm. Chapman and Hall, London, United Kingdom.

Snyder-Mackler, N., Majoros, W.H., Yuan, M.L., Shaver, A.O., Gordon, J.B., Kopp, G.H., Schlebusch, S.A., Wall, J.D., Alberts, S.C., Mukherjee, S., Zhou, X., Tung, J., 2016. Efficient Genome-Wide Sequencing and Low-Coverage Pedigree Analysis from Noninvasively Collected Samples. Genetics 203, 699–714.

Stormont, C., 1967. Contribution of blood typing to dairy science progress. J. Dairy Sci. 50, 253-260.

Strug, L.J., 2018. The evidential statistical paradigm in genetics. Genet. Epidemiol. 42, 590-607.

Strug, L.J., Hodge, S.E., Chiang, T., Pal, D.K., Corey, P.N., Rohde, C., 2010. A pure likelihood approach to the analysis of genetic association data: an alternative to Bayesian and frequentist analysis. Europ. J. Hum. Genet. 18, 933-941.

Vázquez, J.F., Pérez, T., Ureña, F., Gudín, E., Albornoz, J., Domínguez, A., 2004. Practical application of DNA fingerprinting to trace beef. J. Food Protect.67, 972-979.

Vens, M., Schillert, A., König, I.R., Ziegler, A., 2009. Look who is calling: a comparison of genotype calling algorithms. BMC Proceed. 3, S59.

Visscher, P.M., Woolliams, J.A., Smith, D., Williams, J.L., 2002. Estimation of pedigree errors in the UK dairy population using microsatellite markers and the impact on selection. J. Dairy Sci. 85, 2368-2375.

Whalen, A., Gorjanc, G., Hickey, J.M., 2019. Parentage assignment with genotyping-by-sequencing data. J. Anim. Breed. Genet. 136, 102–112.

Weller, J.I., Feldmesser, E., Golik, M., Tager-Cohen, I., Domochovsky, R., Alus, O., Ezra, E., Ron, M., 2004. Factors affecting incorrect paternity assignment in the Israeli Holstein population. J. Dairy Sci. 87, 2627-2640.

Yu, X., Sun, S., 2013. Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinform. 14, 274.

Zan, Y., Payen, T., Lillie, M., Honaker, C.F., Siegel, P.B., Carlborg, Ö., 2019. Genotyping by low-coverage whole-genome sequencing in intercross

pedigrees from outbred founders: a cost-efficient approach. Genet. Sel. Evol. 51, 44.

Table 1 Conditional probability of the offspring's genotype in a biallelic locus (alleles A and B) given the mother's and the alleged father's genotype. Each triad of numbers provides the probability for AA, AB and BB genotypes, respectively.

	Mother's genotype		
Father's genotype	AA	AB	BB
AA	1 / 0 / 0	0.5 / 0.5 / 0	0 / 1 / 0
AB	0.5 / 0.5 / 0	0.25 / 0.5 / 0.25	0 / 0.5 / 0.5
BB	0 / 1 / 0	0 / 0.5 / 0.5	0 / 0 / 1

Table 2 Conditional probability of the offspring's genotype in a biallelic locus (alleles A and B) when only one parent contributes to the paternity test. Each triad of numbers provides the probability for AA, AB and BB genotypes, respectively.

	Parent's genotype		
Offspring's genotype	AA	AB	BB
AA	p(A)[1]	0.5 p(A)	0
AB	1-p(A)	0.5	p(A)
BB	0	0.5 [1 - p(A)]	1 - p(A)

[1]p(A): allelic frequencies of A allele in parents' generation.

Table 3 Mean ± SE of shared polymorphic SNPs among two unrelated individuals when sequenced at different genome coverages with 10,000 base-pair read length.

Genome coverage	Polymorphic SNPs	MAF[1]>0.05 SNPs
0.01X	3,093.2 ± 28.5	1,586.9 ± 17.2
0.05X	77,007.2 ± 75.0	39,887.3 ± 46.4
0.1X	290,845.5 ± 137	151,872.8 ± 86.5
0.5X	4,965,993.9 ± 450.0	2,589,401 ± 329.2

[1]Minimum allele frequency

Fig. 1. Distribution of allele frequencies for the first chromosome of the first simulated population.

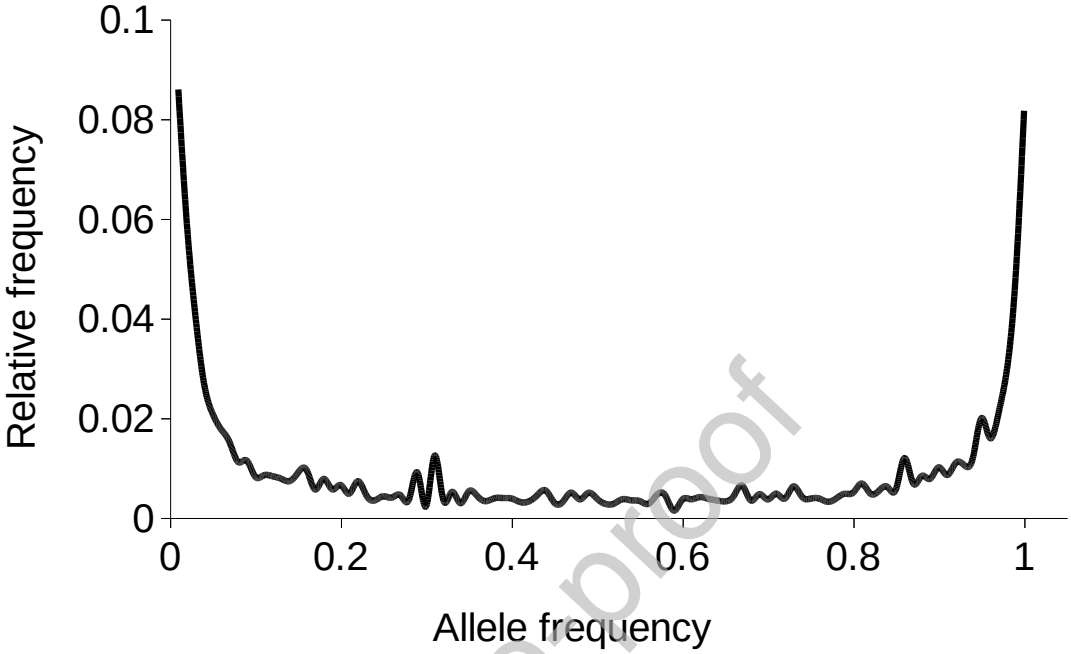


Fig. 2. Average distribution of polymorphic SNPs depending on the number of reads when sequenced at 0.01X (black), 0.05X (blue), 0.1X (red) and 0.5X (white) genome coverage with 10,000 base-pair read length. The whiskers extend to minimum and maximum estimates.

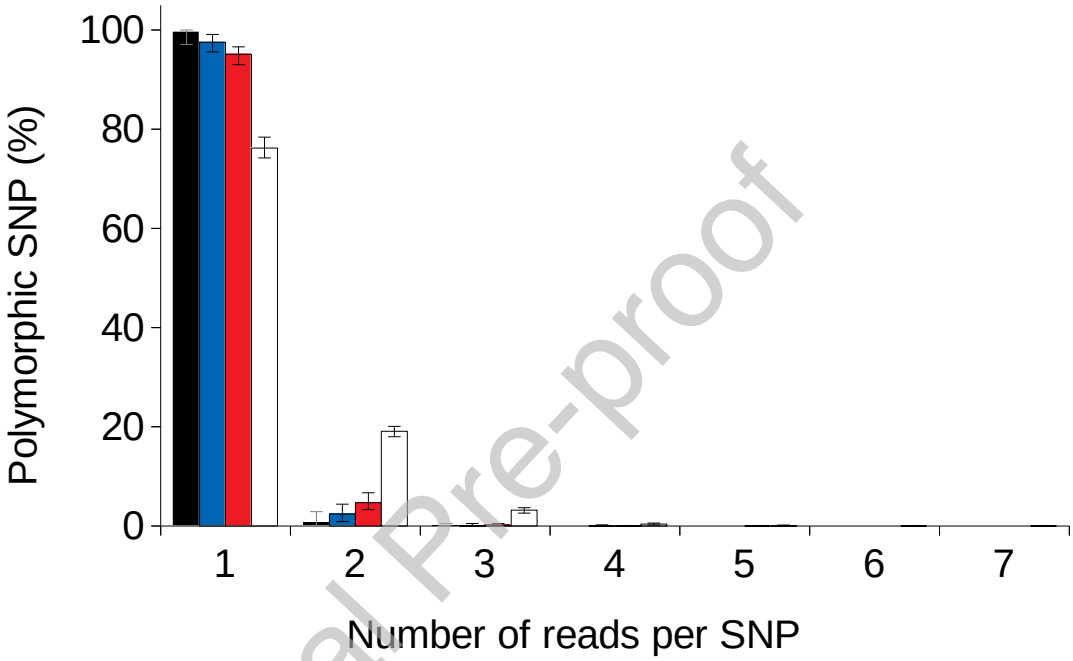


Fig. 3. Shared SNPs among two unrelated individuals both sequenced at 0.01X genome coverage with 100 (red dots), 1,000 (blue dots) and 10,000 base-pair read length (black dots). The X-axis accounts for SNPs with non-zero minimum allele frequency (MAF), whereas Y-axis accounts for SNPs with MAF>0.05.

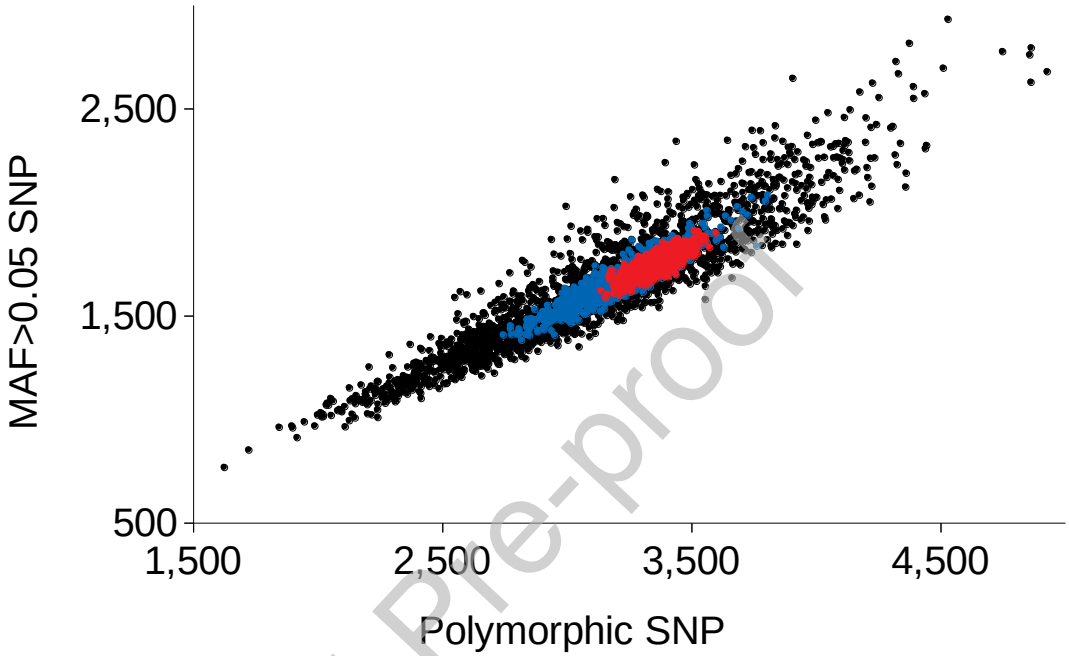


Fig. 4. Distribution of 1,000 traceability tests (black dots) and paternity tests with known (blue dots) and unknown dam (red dots) under four different genome coverage, and assuming true SNP-specific sequencing error rate and true genotype frequencies in parental generation. Traceability tests relied on the likelihood ratio between the null (H_0: same individual) and the alternative hypothesis (H_1: different individuals), and compared each individual against itself (X-axis) and against an unrelated individual (Y-axis). Paternity tests evaluated whether the alleged sire was the true sire (H_0) or an unrelated male of the population (H_1), and where applied on the true sire (X-axis) and on an unrelated male of the population (Y-axis).

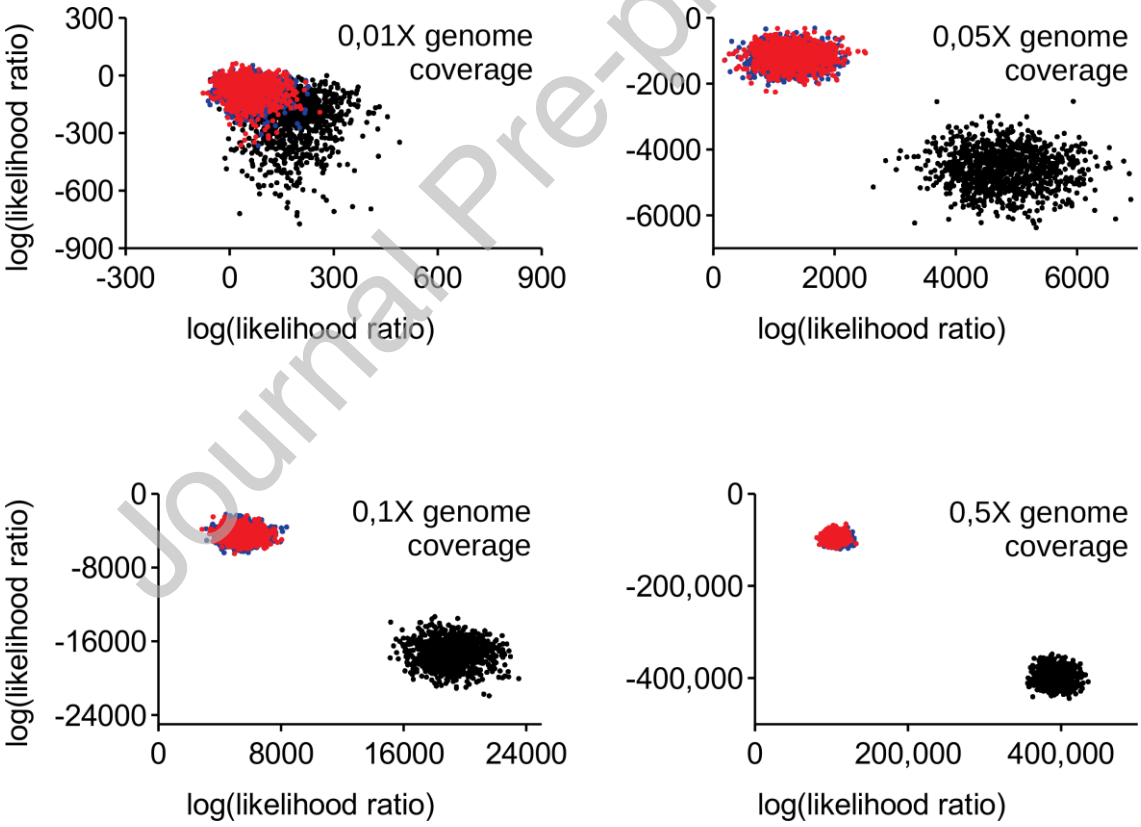
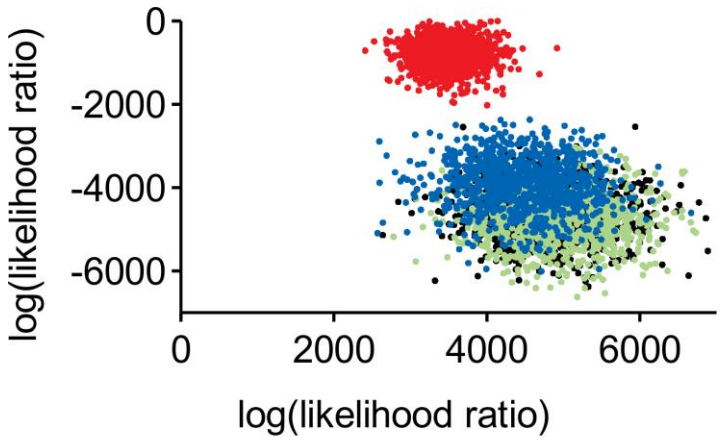
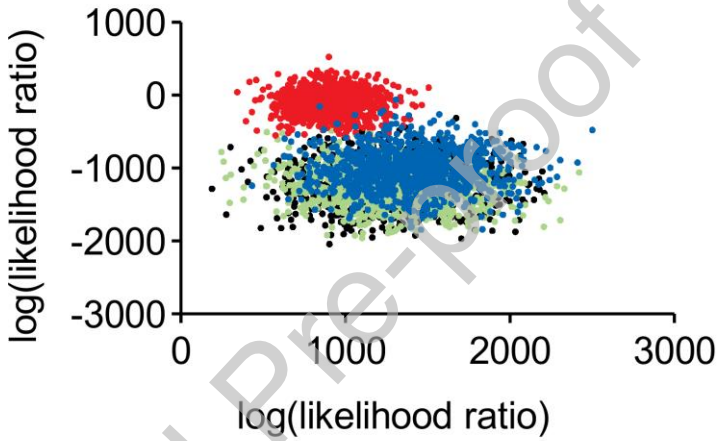


Fig. 5. Distribution of 1,000 traceability tests (upper panel), paternity test with known dam (mid panel) and paternity test with unknown dam (lower panel) under 0.05X genome coverage, 10,000 base-pair read length, and assuming true genotype frequencies in parental generation. Tests assumed true sequencing error rates (black dots), null sequencing error rate (green dots), 1% sequencing error rate (blue dots), and 10% sequencing error rate (red dots). Traceability tests relied on the likelihood ratio between the null (H_0: same individual) and the alternative hypothesis (H_1: different individuals), and compared each individual against itself (X-axis) and against an unrelated individual (Y-axis). Paternity tests evaluated whether the alleged sire was the true sire (H_0) or an unrelated male of the population (H_1), and where applied on the true sire (X-axis) and on an unrelated male of the population (Y-axis).





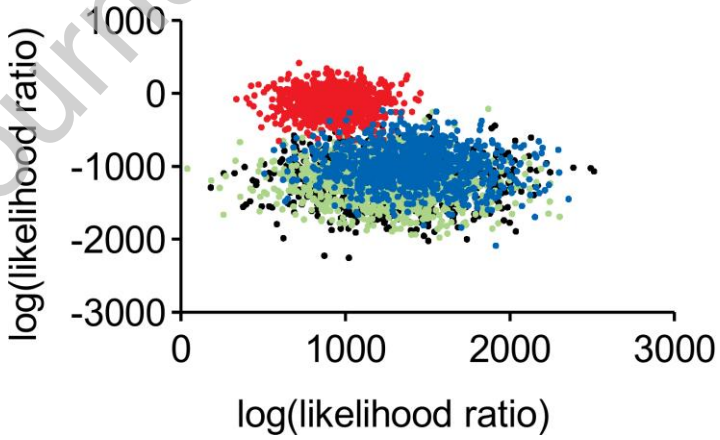


Fig. 6. Distribution of 1,000 traceability tests (upper panel), paternity test with known dam (mid panel) and paternity test with unknown dam (lower panel) under 0.05X genome coverage, 10,000 base-pair read length, and assuming true sequencing error rates per SNPs. Tests assumed true genotyping frequencies (black dots), as well as genotyping frequencies under Hardy-Weinberg equilibrium after sampling the allele frequency from 5 (red dots), 10 (blue dots) and 100 individuals (green dots). Traceability tests relied on the likelihood ratio between the null (H_0: same individual) and the alternative hypothesis (H_1: different individuals), and compared each individual against itself (X-axis) and against an unrelated individual (Y-axis). Paternity tests evaluated whether the alleged sire was the true sire (H_0) or an unrelated male of the population (H_1), and where applied on the true sire (X-axis) and on an unrelated male of the population (Y-axis).